Medical Image Spatial Grounding with Semantic Sampling

Andrew Seohwan Yu[1,2,†], Mohsen Hariri[1,†], Kunio Nakamura[2], Mingrui Yang[2], Xiaojuan Li[1,2], and Vipin Chaudhary[1]

[1] Case Western Reserve University, Cleveland OH 44106, USA
[2] Cleveland Clinic, Cleveland OH 44106, USA

[†] These authors contributed equally. Corresponding author: Vipin Chaudhary (vipin@case.edu).

Abstract. Vision language models (VLMs) have shown significant promise in visual grounding for images as well as videos. In medical imaging research, VLMs represent a bridge between object detection and segmentation, and report understanding and generation. However, spatial grounding of anatomical structures in the three-dimensional space of medical images poses many unique challenges. In this study, we examine image modalities, slice directions, and coordinate systems as differentiating factors for vision components of VLMs, and the use of anatomical, directional, and relational terminology as factors for the language components. We then demonstrate that visual and textual prompting systems such as labels, bounding boxes, and mask overlays have varying effects on the spatial grounding ability of VLMs. To enable measurement and reproducibility, we introduce **MIS-Ground**, a benchmark that comprehensively tests a VLM for vulnerabilities against specific modes of **M**edical **I**mage **S**patial **Ground**ing. We release MIS-Ground to the public at `github.com/asy51/mis-ground`. In addition, we present **MIS-SemSam**, a low-cost, inference-time, and model-agnostic optimization of VLMs that improves their spatial grounding ability with the use of **Sem**antic **Sam**pling. We find that MIS-SemSam improves the accuracy of Qwen3-VL-32B on MIS-Ground by 13.06%.

Keywords: multi-modal large language models · vision language models · 3D medical imaging · visual grounding · spatial grounding

1 Introduction

Merging the embedding spaces of vision and language models through contrastive learning [18] was a revelation that gave rise to the modern vision language model (VLM). Video VLMs were the logical next step, extending these ideas to spatiotemporal data where models must align sequences of frames with natural language. Despite the sheer scale of information in videos relative to images, substantial progress has been made in both video understanding and multimodal reasoning [14].

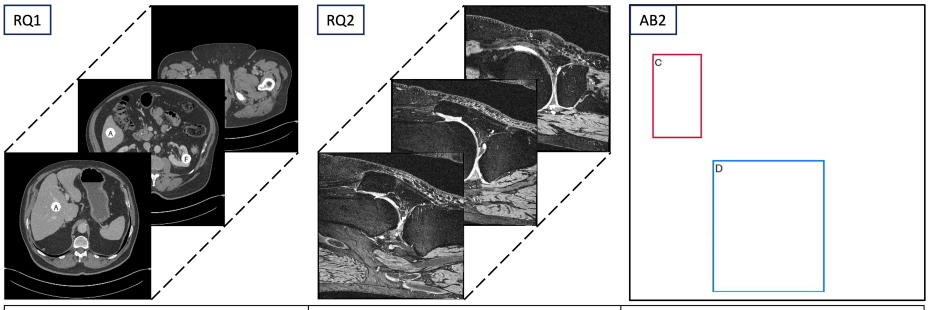
Medical imaging VLMs have been a different matter. As narrow vision models continue to mature in the medical field, medical VLMs are in their early stages of development. Because the available data is limited, models, benchmarks, and training objectives tend to revolve around report generation [8,12] and visual question answering [11,15,13,20], especially disease identification. Though the reported results can be impressive, it is difficult to gauge the practicality of these VLMs in a setting as high-stakes as the medical field. Sambara et al. [19] showed that despite using the correct terminology and making references to anatomical components and lesions, VLMs failed to display reasoning capabilities that were spatially grounded in objects that were actually present in the image in question.

There has been recent work to introduce rigor in spatial grounding to medical VLM training and benchmarking, demonstrating that general-purpose 2D VLMs fail at even simple visual grounding of medical images, due in large part to language-only pretraining that contains ground truths about human anatomy, which may differ from the radiological image in question [23]. Methods for linking anatomical components to language using contrastive alignment in 2D images have also been introduced [10].

To push the current understanding of 3D VLM spatial grounding capabilities, this study proposes **M**edical **I**mage **S**patial **Ground**ing (**MIS-Ground**), a benchmark designed to identify the extent to which video VLMs are able to identify, localize, and differentiate anatomical structures in 3D medical images. MIS-Ground covers a wide variety of language and vision inputs that are likely to occur in medical imaging scenarios. These include different text and visual prompts to align the object of interest to text, multiple views of the same 3D volume, as well as various combinations of question classes and target answers (Figure 1). Because a 3D volume can be serialized as an ordered stack of slices, video-native VLMs offer a practical interface for it; MIS-Ground is nonetheless not video-specific, and we also evaluate 3D-enabled medical VLMs (Section 2.3). To our knowledge, MIS-Ground is the first benchmark to probe 3D anatomical spatial grounding under a controlled, factorial design that jointly varies modality, slice direction, coordinate convention, prompt type, and terminology, whereas prior medical-grounding studies report aggregate failures [19,23] without isolating these factors. This study systematically evaluated five key dimensions of spatial grounding, organized as three research questions (RQ) and two ablations (AB): (RQ1) their capacity for 3D understanding across slice directions; (RQ2) their preference for anatomical over colloquial direction terms; (RQ3) the conditional impact of visual prompts based on orientation and terminology; (AB1) their reliance on pretrained anatomical knowledge priors; and (AB2) their ability to perform abstract spatial reasoning in the absence of medical images.

This study also proposes **M**edical **I**mage **S**patial **Sem**antic **Sam**pling (**MIS-SemSam**), which enhances the reasoning capabilities of the models by employing a sampling strategy based on groups of semantically similar tokens instead of individual ones. This study tested open-weights models Qwen [2,17,4,21,3], MedGemma [20], Molmo [5], and paid model Gemini [6,7] on MIS-Ground; M3D [1] and Med3DVLM [24], two leading 3D medical VLMs, were also tested. With

an overall accuracy of 66.5%, Qwen3-VL-32B with MIS-SemSam demonstrated a strong ability in 3D medical image spatial grounding, despite being trained on videos of natural scenes.



Q: Does the Liver (A) appear before the left kidney? (F) A: True	Q: What is the anatomical spatial relationship between patella bone and the femur bone? A: Anterior	Q: What is the spatial relationship between C and D? A: Above and to the left
Q: What is the slice direction of this image? A: Axial	Q: Is coronal the slice direction of this image? A: False	Q: Is C above D? A: True
Q: What is the name of F? A: Left kidney	Q: Is the patella bone above the femur bone? A: True	Q: Is D to the right of C? A: True
Q: Is A the label of the left kidney? A: False	Q: Is the femur bone posterior to the patella bone? A: True	Q: What is the spatial relationship between D and C? A: Below and to the right

Fig. 1. Sample question-answer pairs from MIS-Ground. For each 2D or 3D vision input (top), multiple questions are generated (bottom). Video VLMs display the ability to discern anatomical structure relationships across slices (RQ1); they prefer anatomical direction terms like anterior (RQ2); and they are capable of abstract spatial reasoning (AB2). Note: the CT image is shown in axial RAS standard viewing orientation, whereas the MRI is shown in sagittal RAS storage mode.

2 Methods

MIS-Ground was designed to systematically test the spatial grounding capabilities of VLMs against a variety of scenarios in medical imaging. This requires a careful curation of the dataset and data pipeline to query the VLMs effectively and efficiently. These integrations are included as part of MIS-Ground. VLMs with and without MIS-SemSam are treated as separate models and thus can be tested on MIS-Ground individually.

2.1 Dataset Preparation

Knee MRI from the Osteoarthritis Initiative (OAI) [16] and torso CT from TotalSegmentator (CT dataset) [22] were used in this study. With access to the MRI dataset under the appropriate data use agreement, 209 double echo steady state (DESS) MR images of the knee joint were collected. For each of these scans, 3D

bounding boxes for 10 anatomical components in the knee were gathered from 3DReasonKnee-Bench [19], an open license (CC BY 4.0) benchmark dataset. These components included bone, cartilage, ligaments, and other regions in the knee joint that are associated with osteoarthritis. For the CT dataset, 1228 full-dose axial scans of the torso were collected partly under an open license (CC BY 4.0) and a non-commercial license from TotalSegmentator [22]. Under the same licenses, up to 117 matching segmentation masks of anatomical components were collected for each scan. Prioritizing visibility and ease of prompting, 67 of the larger and more commonly named components were selected for this study, and scans with fewer than 200 or greater than 450 slices were omitted. For preprocessing, MRI and CT datasets underwent percentile-based and soft-tissue windowing, respectively. Multi-planar reconstruction with trilinear interpolation was used to render the scans in alternative slice directions. Finally, the scans were placed in one of two orientations: RAS (right-anterior-superior) coordinate system storage (where $[0, 0, 0]$ is the RAS-most point) and standard viewing orientations.

2.2 Benchmark Curation

MIS-Ground consists primarily of questions evaluating the relative positions of anatomical structures. Questions were defined by four parameters: (1) Visual prompts were derived from ground-truth annotations, including masks, bounding boxes, and centroids (points) for the CT dataset, and bounding boxes for the MRI dataset. Prompts were uniquely colored and/or labeled with a letter A-F. (2) Text prompts included names of anatomical structures and/or the colors or letters of the visual labels. (3) Target type determined the target answers of the generated queries: structure names, labels, or spatial relationships, classified as either anatomical (e.g. superior) or colloquial (above). (4) Question type formatted the queries as open-ended, closed (True), or closed-inverted (False). Ablations included omitting visual inputs entirely (text-only) or omitting the medical image itself (visual prompts on a white background). To ensure broad coverage, pairs of anatomical structures were randomly selected for each parameter combination, yielding an average of 683 and 2,757 questions per MRI and CT scan, respectively.

2.3 VLM Evaluation

A variety of video-native open-weights VLMs were benchmarked on MIS-Ground, including Alibaba's Qwen2.5-VL-Instruct (3B, 7B, 32B, 72B) [4] and Qwen3-VL-Instruct (2B, 4B, 8B, 32B) [2]; and AllenAI's Molmo 2 (4B, 8B) [5]. In addition, 3D-enabled medical VLMs were tested, including Google's MedGemma (4B, 27B) and MedGemma 1.5 (4B) (based on Gemma 3) [20]; M3D (based on Llama 2 7B) [1]; and Med3DVLM (based on Qwen2.5 7B) [24]. For individual model performances, see Figure 2. Additionally, closed-source models, Google's Gemini models (2.5 Flash, 2.5 Flash Lite, 3 Flash) were benchmarked as well. To

ensure fairness in benchmark comparisons and to promote reproducibility, inference configurations were tuned for the same sampling configuration: reasoning with max new tokens=8,912 and temperature=0.5. To report performance robustly under limited trials, we report the Bayesian credible interval [9]. Due to model design, inputs to M3D and Med3DVLM were resized to $[32, 256, 256]$ and $[128, 256, 256]$ respectively with trilinear interpolation, and may have resulted in reduced performance.

2.4 Semantic Sampling

As discussed in Section 1, many failures in medical image spatial grounding are not due to a complete lack of visual recognition, but rather to **language-side brittleness**: small decoding perturbations can flip directional terms (e.g., *superior* vs. *inferior*), select awkward morphological variants, or drift into semantically adjacent but incorrect anatomical phrasing. This effect becomes more pronounced under long-form reasoning (chain of thought), where a single locally plausible token can derail the subsequent trajectory. To mitigate this issue at inference time without additional training, we introduce **MIS-SemSam**, an adaptation of Semantic Sampling to medical VLM spatial grounding.

In Semantic Sampling, instead of selecting the next token purely by its own probability, Semantic Sampling prefers tokens that are supported by probability mass in their local semantic neighborhood in token-embedding space. Intuitively, if the model is unsure among several semantically similar realizations of the same concept (common in directional and anatomical terminology), aggregating neighborhood mass stabilizes decoding toward semantically coherent regions of the vocabulary.

Offline semantic neighborhoods. Let the language component of a VLM have an input embedding matrix $E \in \mathbb{R}^{V_{\text{emb}} \times d}$ and a tokenizer vocabulary V_{tok}. Following Semantic Sampling decoding, we partition token ids into *content tokens* C (ordinary text) and *non-content tokens* U (special, added or control tokens and any reserved embedding rows not produced by the tokenizer). This distinction is particularly important for VLMs because modality and chat-template markers (e.g., image delimiters, role tokens) behave as control symbols and can form embedding space "hubs" that contaminate nearest-neighbor structure.

We therefore construct neighborhoods *only over content tokens*. Let $E_C = E[C]$ denote the content-token embeddings. We row-normalize

$$\hat{e}_i \;=\; \frac{E_C[i]}{\|E_C[i]\|_2 + \epsilon}, \quad i \in C, \tag{1}$$

so that cosine similarity is $\cos(i, j) = \hat{e}_i^\top \hat{e}_j$. For each $i \in C$, we precompute its K nearest neighbors in C under cosine similarity (including i itself) and store two lookup tables:

$$S_{\text{tid}} \in \mathbb{N}^{|C| \times K}, \quad S_{\text{val}} \in \mathbb{R}^{|C| \times K}, \tag{2}$$

where $S_{\text{tid}}[i, k]$ is the k-th neighbor id and $S_{\text{val}}[i, k]$ its cosine similarity. These tables are computed once per embedding space and reused across all MIS-Ground

queries; moreover, when multiple checkpoints share the same tokenizer indexing, the same neighbor *ids* can be reused as in Semantic Sampling.

Inference-time semantic rescoring. At decoding step t, the VLM outputs logits $\ell_t \in \mathbb{R}^{V_{\text{emb}}}$ over the next text token, which induce (temperature-scaled) probabilities

$$p_t(v) = \frac{\exp(\ell_t(v)/T)}{\sum_{v'} \exp(\ell_t(v')/T)}. \tag{3}$$

We first apply any standard truncation filter $\textsc{Filter}(\cdot)$ (e.g., top-M, top-p) to obtain a candidate set $I_t \subseteq V_{\text{emb}}$. Because semantic neighborhoods are defined only for content tokens, if $I_t \cap U \neq \emptyset$, we skip MIS-SemSam for this step and defer to the model's default decoding rule (greedy for T=0, or the configured sampler for T>0). Otherwise, $I_t \subseteq C$ and we compute a semantic score for each candidate $c \in I_t$ by aggregating probability mass over its neighborhood:

$$\text{Score}_t(c) \;=\; \sum_{k \in \mathcal{K}(c)} w_{c,k}\, p_t\big(S_{\text{tid}}[c,k]\big), \qquad w_{c,k} = \max\big(0,\, S_{\text{val}}[c,k]\big), \tag{4}$$

where $\mathcal{K}(c)$ denotes the kept neighbor indices (either the top-K' neighbors, $K' \leq K$, or a similarity-threshold prefix), and negative cosine values are clamped to zero to prevent cancellation.

Finally, the next token is chosen from I_t using either: (i) deterministic selection $x_{t+1} = \arg\max_{c \in I_t} \text{Score}_t(c)$ (used to maximize reproducibility), or (ii) stochastic selection by sampling from a softmax over Score_t (used in chain of thought settings to preserve controlled exploration). MIS-SemSam requires no additional forward passes and adds only $O(|I_t| \cdot |\mathcal{K}(c)|)$ table lookups per token.

MIS-SemSam is applied as a drop-in replacement for the final "pick next token" step in generation, leaving the vision encoder, prompting strategy, and truncation filter unchanged. In our experiments, we treat a model with MIS-SemSam enabled as a separate inference configuration (Section 2.3), enabling apples-to-apples comparisons between default decoding and semantic-neighborhood-supported decoding on identical MIS-Ground questions. The reported improvement is therefore a paired comparison by construction: the model, prompts, images, questions, and inference settings are held fixed, and only the final token-selection rule changes. We use *model-agnostic* in the algorithmic sense—MIS-SemSam applies to any VLM that exposes next-token logits and input embeddings, and requires neither training nor additional forward passes—rather than as a claim of broad empirical validation; our strongest empirical evidence is on the Qwen3-VL family, while MIS-Ground itself spans many model families.

3 Results

A total of 1160 3D volumes, 2320 2D slices, and 33864 questions were generated across both CT and MRI subsets, which contained 67 and 10 labeled anatomical components per sample respectively. In an effort to balance the dataset for each anatomical component, the CT subset was sampled more and accounted

for 84% of the entire dataset. Consequently, the aggregate score is CT-weighted, and results should be read at the benchmark level rather than as uniform cross-modality or clinical-readiness claims. Sample question-answer pairs are shown in Figure 1. Model response was instructed to be placed at the end in special tags, after all the reasoning tokens. If the tags were missing, the question was omitted from analysis; this omission reflects instruction-following rather than spatial correctness, and predominantly affected the smallest models and the 3D medical VLMs M3D and Med3DVLM, whose outputs were frequently malformed.

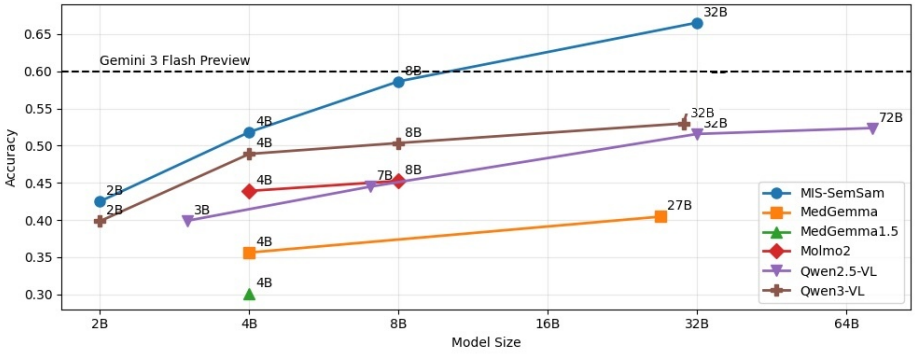


Fig. 2. Overall accuracy on the MIS-Ground benchmark, by model family and size. Note that Med3DVLM (0.0%) and M3D (15.9%) were tested but omitted from this graph. The MIS-SemSam family is derived from the Qwen3-VL family.

Model version and size was strongly correlated to overall performance, with Qwen3-VL-32B outperforming all other open-weights models by a comfortable margin. The overview of these results are found in Figure 2. M3D and Med3DVLM both failed to give valid responses, often giving long-winded explanations for unrelated concepts, often in non-English characters. These failures likely reflect aggressive input resizing and rigid response formatting (Section 2.3) rather than an absence of spatial capability, and should not be read as a general verdict on 3D medical VLMs. By contrast, the closed-source Gemini models served as a strong reference point, underscoring that MIS-Ground is useful beyond our own method. Medgemma did manage to follow the prompt for the most part, though it still gave poor responses. On the other hand, Qwen and Molmo models had much more success, though the very small models (2B, 3B and 4B) still struggled to give valid responses. Large Qwen2.5 models, especially 72B, underperformed relative to its size, which confirmed Qwen's iterative improvements between versions. Otherwise, no strongly discernible patterns emerged regarding model class and specific performances. As the best performing open-weights model, MIS-SemSam (Qwen3-VL-32B with semantic sampling) was used to evaluate specific spatial grounding capabilities across the following research questions and ablations.

(RQ1) 3D Understanding: MIS-SemSam effectively processed 3D medical images, accurately predicting slice direction in 77.6% of cases. The model successfully determined spatial relationships across slices (68.0% accuracy) nearly as well as in-plane relationships (71.7%). Overall performance on 3D images (56.9%) remained competitive with 2D images (58.8%).

(RQ2) Direction Terms: The model demonstrated a strong preference for anatomical directional terms over colloquial terms. In standard viewing orientations, anatomical terms outperformed colloquial terms (69.4% vs. 57.8%). However, the model shifted its preference to colloquial terms (59.85% vs. 50.2%) in RAS storage mode, where standard medical priors conflict with the visual presentation.

(RQ3) Visual Prompts: The impact of visual prompts varied depending on orientation and terminology. In standard RAS viewing orientation, adding points, bounding boxes, or masks yielded only modest improvements (+2.97%, +2.33%, and +2.52% respectively) over the 57.6% baseline. For RAS storage mode utilizing anatomical terms, visual prompts severely degraded a strong 75.3% baseline performance (-13.33% for points, -9.22% for boxes) due to conflicting information. Conversely, when forced to use colloquial terms, visual prompts significantly aided the model, improving a weak 45.1% baseline by up to +8.23%.

(AB1) Anatomy Prior: The model relied heavily on pretrained anatomical knowledge. In text-only ablations (no image), it achieved 69.3% accuracy on anatomical relationships (compared to 74.0% with the image). This prior did not extend to colloquial terms, where text-only accuracy dropped to 40.4%.

(AB2) Abstract Reasoning: The model demonstrated capable abstract spatial reasoning. When tested on visual prompts placed against a blank background without medical scans, the model still accurately determined spatial relationships with 64.2% (points) and 60.4% (boxes) accuracy.

4 Conclusions

With MIS-Ground, VLMs demonstrated measurable capabilities in spatial grounding of anatomical structures in medical imaging. This is in contrast to previous studies, which reported apparent random guessing and failure to follow instructions [19,23]. This is due to careful considerations made in the construction of MIS-Ground, which prioritized the balance and diversity of image modality, visual prompts, text prompts, target answer types, question types, image orientation, slice direction, 2D and 3D, terminology use, and ablations. MIS-SemSam also represents an exploration into the optimization of the VLMs in medical

image spatial grounding: an inference-time method with negligible costs and significant performance improvements.

More improvements and applications of VLMs lie ahead. For one, fine-tuning is a crucial task for a domain as highly specialized as medical imaging. Unfortunately, many current such paradigms seem to overfit the task of existing medical visual question answering benchmarks and fail to translate to general capabilities like spatial grounding. New developments in such benchmarks and training datasets are necessary for this field to continue to grow. Also, spatial grounding of anatomical structures is only a partial coverage of what is needed in clinical practice, such as report generation. There is still a lack of alignment between the model capabilities and practical use. To help the community build toward these goals, we release our benchmark MIS-Ground to the public at `github.com/asy51/mis-ground`.

Acknowledgments. This research was supported in part by NSF awards 2117439 and 2320952.

Disclosure of Interests. The authors declare they have no competing interests.

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., et al.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Bai, S., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Clark, C., Zhang, J., Ma, Z., Park, J.S., Salehi, M., Tripathi, R., Lee, S., Ren, Z., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611 (2026)
6. Gemini Team: Gemini 2.5: Pushing the frontier with advanced reasoning. arXiv preprint arXiv:2507.06261 (2025)
7. Google DeepMind: Gemini 3 flash (model documentation). `https://deepmind.google/models/gemini/` (2026)
8. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., et al.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography. arXiv preprint arXiv:2403.17834 (2024)
9. Hariri, M., Samandar, A., Hinczewski, M., Chaudhary, V.: Don't pass@k: A bayesian framework for large language model evaluation. In: The Fourteenth International Conference on Learning Representations (2026), `https://openreview.net/forum?id=PTXi3Ef4sT`
10. Hashmi, A.U.R., Saeed, N., Lippert, C.: Anatomix, an anatomy-aware grounded multimodal large language model for chest x-ray interpretation. arXiv preprint arXiv:2601.03191 (2026)

11. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
12. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
13. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018)
14. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
15. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654. IEEE (2021)
16. Peterfy, C.G., Schneider, E., Nevitt, M.: The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. Osteoarthritis and Cartilage **16**(12), 1433–1441 (2008). https://doi.org/10.1016/j.joca.2008.06.016
17. Qwen Team: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
19. Sambara, S., Kim, S.E., Zhang, X., Luo, L., Johri, S., Baharoon, M., Ro, D.H., Rajpurkar, P.: 3dreasonknee: Advancing grounded reasoning in medical vision language models. In: Biocomputing 2026: Proceedings of the Pacific Symposium. pp. 99–113. World Scientific (2025)
20. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
21. Wang, P., et al.: Qwen2-vl technical report. arXiv preprint arXiv:2409.12191 (2024)
22. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
23. Wolf, D., Hillenhagen, H., Taskin, B., Bäuerle, A., Beer, M., Götz, M., Ropinski, T.: Your other left! vision-language models fail to identify relative positions in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 691–701. Springer (2025)
24. Xin, Y., Ates, G.C., Gong, K., Shao, W.: Med3dvlm: An efficient vision-language model for 3d medical image analysis. IEEE Journal of Biomedical and Health Informatics (2025)